METHOD FOR COMPARING SIGNAL ARRAYS IN DIGITAL IMAGES

FIELD OF THE INVENTION

This U.S. patent application claims priority from Israel Patent Application No. 141151 of January 29, 2001. The invention relates to methods of comparing the intensity of two signal arrays in digital images, for example digital images of a spot in a one- or two-dimensional electrophoresis pattern or a DNA chip.

BACKGROUND OF THE INVENTION

A digital image may be considered to be an array of signals, where each pixel in the image produces a visible signal of a particular intensity. It is often of interest to compare two such signal arrays. For example, two protein mixtures can be separated by one of various separation techniques to produce two one- or two-dimensional separation patterns. A digital image of a spot in each pattern, corresponding to the same protein could be compared in order to compare the amount of the protein present in each mixture. As another example, a DNA chip having attached to it various oligonucleotide targets is incubated in the presence of probe oligonucleotides from two sources. The two probe species are differently labeled, so that each probe species produces a visible signal that is distinguishable from that of the other species. For example, one probe species may be labeled with a fluorescent dye that produces a red signal while the other probe species is labeled with a fluorescent dye that produces a green signal. A digital image of the red signal could then be compared with a digital image of the green signal in order to compare the amount of oligonucleotides binding to the chip in the two sources.

One well-known method for comparing the signal arrays in two digital images involves calculating the total intensity in each image and then calculating

the ratio of these two intensities. Another method is to determine the maximum intensity in each image and to calculate the ratio of the two maximal intensities.

DESCRIPTION OF THE INVENTION

The present invention provides a method for comparing two visual signal arrays. A signal array may be, for example, a digital image of a stained spot in a one-or two- dimensional separation pattern such as produced by electrophoresis. A signal array may also be a digital image of a region of a DNA chip that has been incubated with labeled probes that produce a visible signal. The two arrays to be compared may be physically separated from one another or superimposed upon one another.

In one embodiment of the invention, the two signal arrays to be compared are superimposed upon one another. The two arrays may be, for example, a single digital image of a region on a DNA chip that was simultaneously incubated with nucleic acid probes from two different sources, where the probes from each source are labeled with a marker producing a distinct visible signal. For example, the probes from one source may be labeled with a fluorescent label producing a red signal, and the probes from the other source labeled with a label producing a green signal. In this case, the red and green signal arrays in the digital image are superimposed upon one another, and are to be compared by the method of the invention.

When the two arrays are superimposed upon one another, each pixel x_i in the superimposition is described by an ordered pair of numbers $(I_1(x_i), I_2(x_i))$ where $I_1(x_i)$ is the intensity of the signal of the pixel x_i in the first array, and $I_2(x_i)$ is the intensity of the signal of the pixel x_i in the second array. A linear regression analysis is applied to the points $(I_1(x_i), I_2(x_i))$. Within the context of the present invention, the term *"linear regression"* is used to include any method in which a linear fit is found for a set of points, for example, a least squares fit of the points to a line, as is known in the art. This also includes methods involving a filtering step in which points are deleted from the set of points prior to determining the

linear fit. In accordance with the invention, the two arrays are compared by means of the slope of the line produced by the linear regression analysis.

In another embodiment of the invention, two signal arrays are compared that are not superimposed upon one another. The two patterns may be, for example, digital images of spots in different one- or two- dimensional separation patterns such as produced by electrophoresis. The two arrays are first put into register with each other. Registration of the two patterns is described by means of a transformation **T** that maps a pixel x_i in the first pattern to a pixel $\mathbf{T}(x_i)$ in the second pattern. Methods for obtaining registration transformations are disclosed, for example, in Israel Patent Application 133562 Two arrays in register with each other under the transformation **T** are compared in accordance with the invention as follows. For each pixel x_i in the first array, an ordered pair of numbers $(I(x_i), I(\mathbf{T}(x_i))$ is generated where $I(x_i)$ is the intensity of the signal of a pixel x_i in the first array and $I(\mathbf{T}(x_i))$ is the intensity of the pixel $\mathbf{T}(x_i)$ in the second pattern that is in register with the pixel x_i. A linear regression analysis is applied to the points $(I(x_i), I(\mathbf{T}(x_i))$. In accordance with the invention, the two arrays are compared by means of the slope of the regression line produced by the linear regression analysis.

The invention may be used for the determination of differential gene expression. In this application, each of the signal arrays to be compared represents the level of expression of a particular gene. Typically, but not necessarily, the two arrays represent the level of the gene expression under different conditions. The invention may also be used for the determination of differential protein expression. In this application, each of the signal arrays to be compared represents the amount of a particular protein present in a sample.

BRIEF DESCRIPTION OF THE DRAWINGS

In order to understand the invention and to see how it may be carried out in practice, a preferred embodiment will now be described, by way of non-limiting example only, with reference to the accompanying drawings, in which:

Fig. 1 is a plot of the ordered pairs $(I_1(x), I_2(x_i))$ where $I_1(x_i)$ is the intensity of a signal produced by a first DNA probe species in the pixel x_i, $I_2(x_i)$ is the intensity of a signal produced by a second DNA probe species in the pixel x_i, the DNA probes being bound to DNA targets on a DNA chip;

Fig. 2 shows two two-dimensional separation patterns;

Fig. 3 shows a enlargement of first and second spots from the first and second separation patterns, respectively, of Fig. 2, and

Fig. 4 shows a plot of the points $(I(x_i), T(I(x_i)))$, where $I(x_i)$ is in the intensity of a pixel x_i in the first spot of Fig. 3 and $I(\mathbf{T}(x_i))$ is the intensity of a pixel $\mathbf{T}(x_i)$ in the second spot that is in register with the first spot under a transformation T.

EXAMPLES

Example 1 Two superimposed spots

A DNA chip having DNA targets bound on it was incubated in the presence of a sample containing first and second DNA probe species, where each probe species was labeled with a label producing a distinct visible signal. Each of the first and second probe species bound to a particular target on the chip thus produces a distinct signal array in a region of the chip where the target is located. For a pixel x_i, the intensity of the two signal arrays is represented by an ordered pair of numbers $(I_1(x_i), I_2(x_i))$ where $I_1(x_i)$ is the intensity of the signal produced by the first probe species in the pixel x_i and $I_2(x_i)$ is the intensity of the signal produced by the second probe species in the pixel x_i. Fig. 1 shows a plot of the ordered pairs $(I_1(x_i), I_2(x_i))$. A linear regression analysis was applied to the points $(I_1(x_i), I_2(x_i))$ that produced the best linear fit **200** to the points. The slope of the line **200** was found to be 1.48, indicating that a probes of the second species binding to a particular target on the chip were present in the sample at an abundance of about 1.48 times that of probes of the first species binding to the same target. The two spots are compared by means of the slope of the line **200.**

Example 2 Separated arrays

Two samples containing proteins are separated to produce a pair of two-dimensional separation patterns. Fig. 2 shows a representation of two

two-dimensional separations patterns **305** and **310**. A spot **315** in the first pattern **305** is to be compared with a spot **320** in the second pattern **310**. Fig. 3 shows enlargements of the spots **315** and **320**, divided into pixels. The pixels in each spot form a signal array. Each pixel in the spot **315**, for example, the pixel **325** has an associated intensity $I(x_i)$. Similarly, each pixel y_i in the spot **320**, for example the spot **330**, has an associated intensity $I(y_i)$. A mapping **T** is found that maps each of a plurality of pixels in the spot **315** to a different pixel in the spot **320**. For example, the pixel **325** may be mapped into the pixel **330**.

If the two spots **315** and **320** consist of the same number of pixels, then the mapping **T** may be obtained by first putting the entire patterns **305** and **310** into register with each other. The patterns **305** and **310** are put in register with one another by means of a transformation **T** that maps each pixel x_i in the pattern **305**, for example the pixel **330** to a pixel $\mathbf{T}(x_i)$ in the pattern **310**. A transformation that puts the two patterns into register with each other may be found, for example, as disclosed in Israel Patent Application No. 133562. The restriction of the transformation **T** to the spot **315** maps pixels in the spot **315** to pixels in the spot **320**.

Another method that may be used to put the spots **315** and **320** into register with each other when the two spots consist of about the same number of pixels is to arrange the pixels in each spot in order of decreasing intensity. The mapping **T** is then defined that maps the nth pixel in the arrangement of the pixels of the spot **315** with the nth spot in the arrangement of the pixels of the spot **320**.

When the two spots **315** and **320** consist of about the same number of pixels, and the mapping **T** has been defined, pairs of numbers are $(I(x_i), I(\mathbf{T}(x_i)))$ formed where $I(x_i)$ is in the intensity of a pixel x_i in the pattern **105** and $I(\mathbf{T}(x_i))$ is the intensity of the pixel $\mathbf{T}(x_i)$ in the pattern **115** that is in register with x_i under the transformation **T**. Fig. 4 shows a plot of the points $(I(x_i),\mathbf{T}(I(x_i)))$. A linear regression analysis is applied to the points that produces the best linear fit **400** to the points. The slope of the linear fit **400** is found to be 4.8 indicating that the spot

320 contains about 4.8 as much protein as is present in the spot **315**. The two spots are compared by means of the slope of the line **400**.

If, say, the spot **315** consists of substantially more pixels than the spot **320**, the following method may be used to put a plurality of the pixels of the spot **315** into register with pixels in the spot **320**. The pixels in each spot are arranged in order of decreasing intensity. A predetermined fraction r_1 of the pixels in the spot **315** are then deleted from the arrangement of the pixels of that spot, to produce a provisional arrangement of the pixels of that spot. A predetermined fraction r_2 of the pixels in the spot **320** are then deleted from the arrangement of the pixels of that spot, to produce a provisional arrangement of the pixels of that spot. r_1 and r_2 are selected so that the two provisional arrangements consist of about the same number of pixels. Preferably, the pixels deleted to form the provisional arrangements are substantially uniformly distributed in each of the initial arrangements. Thus, about every $1/r_1$-th pixel is removed from the initial sequence of pixels from the spot **315** and about every $1/r_2$-th pixel is removed from the initial sequence of pixels from the spot **320**. A transformation **T'** is then defined that maps the nth pixel in the provisional arrangement of the pixels of the spot **315** with the nth spot in the provisional arrangement of the spot **320**.

Pairs of numbers are $(I(x), I(\mathbf{T}'(x_i)))$ formed where $I(x_i)$ is in the intensity of a pixel x_i in the pattern **105** and $I(\mathbf{T}'(x_i))$ is the intensity of the pixel $\mathbf{T}'(x_i)$ in the pattern **115** that is in register with x under the transformation **T'**. Fig. 5 shows a plot of the points $(I(x_i),\mathbf{T}'(I(x_i)))$. A linear regression analysis is applied to the points that produces the best linear fit **500** to the points. The slope of the linear fit **500** is multiplied by r_2/r_1 to compensate for the deletion of points from the two spot arrangements.

It will also be understood that the system according to the invention may be a suitably programmed computer. Likewise, the invention contemplates a computer program being readable by a computer for executing the method of the invention. The invention further contemplates a machine-readable memory tangibly

embodying a program of instructions executable by the machine for executing the method of the invention.